Ligand biological activity predicted by cleaning positive and negative chemical correlations

Alpha A. Lee[a,1], **Qingyi Yang**[b], **Asser Bassyouni**[c], **Christopher R. Butler**[b], **Xinjun Hou**[b], **Stephen Jenkinson**[c], **and David A. Price**[b]

[a]Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, United Kingdom; [b]Pfizer Inc. Medicine Design, 1 Portland Street, Cambridge, Massachusetts USA 02139; [c]Drug Safety Research and Development, Pfizer Inc., La Jolla, CA 92121

This manuscript was compiled on February 28, 2019

Predicting ligand biological activity is a key challenge in drug discovery. Ligand-based statistical approaches are often hampered by noise due to undersampling: the number of molecules known to be active or inactive is vastly less than the number of possible chemical features that might determine binding. We derive a statistical framework inspired by random matrix theory and combine the framework with high quality negative data to discover important chemical differences between active and inactive molecules by disentangling undersampling noise. Our model outperforms standard benchmarks when tested against a set of novel and challenging retrospective tests. We prospectively apply our model to the human muscarinic acetylcholine receptor M1, finding 4 experimentally-confirmed novel agonists that are chemically dissimilar to all known ligands. The hit rate of our model is significantly higher than the state of the art. Our model can be interpreted and visualized to offer novel chemical insights about the molecular motifs that are synergistic or antagonistic to M1 agonism, which we have prospectively experimentally verified.

Finding novel hits to a target receptor is an important initial step in the long process of drug discovery. Although biochemical assays are increasingly high throughput, an experiment-only strategy that attempts to screen chemical space exhaustively remains intractable. To accelerate drug discovery, computer-aided virtual screening strategies have been developed in the literature over the last decades [1–5]. Structure-based approaches require knowledge of the receptor structure and the binding site, and predict the binding free energy by modelling protein-ligand interactions [6–8]. However, determining the receptor structure and parametrising protein-ligand interactions are often challenging, and notwithstanding those challenges it is still computationally intensive to compute the protein-ligand binding free energy [9, 10].

Ligand-based methods sidestep the challenges of structure-based approaches and only require a set of molecules that are known to be active against a particular receptor or triggers a particular phenotype [11, 12]. Those methods are built on the hypotheses that the receptor binding site recognises a specific set of chemical motifs in a molecule, and those motifs can be uncovered from chemical motifs that are shared between the known active molecules. Therefore an unknown molecule is likely to be active if it also contains those common chemical motifs, and inactive otherwise. Those chemical motifs characterise the protein binding site and proteins can be related based on the chemical similarity of their ligands [13, 14].

However, regardless of how one defines chemical motifs – common strategies include using pharmacological intuitions [16–18], enumerating all linear or circular fragments below a certain size around every atom [18, 19], or unsupervised learning [20] – a molecule has many motifs but only a few are important for biological activity of a particular target. Unless one fortuitously knows a priori which motifs are important for a specific receptor and eliminate the other "nuisance" motifs, with a finite amount of data the nuisance motifs could drown out the motifs that are actually important for binding simply by chance. This problem is all the more challenging as it is often the confluence of different motifs, rather than a single motif, that drives binding, yet correlations are known to be especially sensitive to noise due to undersampling [21–23]. Pioneering advances in machine learning that infer the optimal representation of molecules directly from data, e.g. refs [24–27], do not resolve this undersampling problem as the available data is usually significantly less than the number of parameters in the model.

In this paper, we show that removing the noise arising from statistical undersampling – having not enough samples compared to the number of motifs – is needed to reveal chemical differences between active (positive) and inactive (negative) molecules and identify important chemical motifs that determines activity. We develop a statistical method based on random matrix theory and use our model to prospectively discover experimentally-confirmed novel agonists of human muscarinic acetylcholine receptor M1 that are chemically dissimilar to known ligands. Our model also compares favourably with the prior art on retrospective benchmark tests. Importantly, we

Significance Statement

Predicting ligand biological activity is a key challenge in drug discovery. Although there is an increasing amount of activity data, a data-driven approach needs to overcome the challenge that the number of molecules known to be active or inactive is vastly less than the number of possible chemical features that might determine binding. We develop a framework using random matrix theory that discovers important chemical features by disentangling undersampling noise. This method is used to prospectively discover 4 experimentally-confirmed novel agonists of the human muscarinic acetylcholine receptor M1, a target for diseases such as Alzheimer's disease and schizophrenia. Crucially, our method is interpretable, and yields novel and prospectively validated chemical insights on the binding modes of the M1 receptor.

Conceptualization, A.A.L., Q.Y., C.R.B, X.H., D.P.; Methodology, A.A.L., Q.Y., C.R.B, X.H., D.P.; Algorithm development, A.A.L, Experiment, A.B., S.J. ; Writing, AAL.

Q.Y., A.B., C.R.B., X.H., S.J. and D.A.P. are current employees of Pfizer. Structure highlighted in red in Fig 2C is exemplified in the patent WO/2013/072705.

[1] To whom correspondence should be addressed. E-mail: aal44@cam.ac.uk (AAL)

Target	RMD (Eq. 3)	Random matrix	Naive Bayes	Tanimoto	SVM	Graph Convolutions
MOR1	0.99 ± 0.001	0.91 ± 0.003	0.95 ± 0.003	0.91 ± 0.005	0.70 ± 0.01	0.93 ± 0.007
5-HT2B	0.93 ± 0.007	0.82 ± 0.005	0.85 ± 0.01	0.85 ± 0.008	0.67 ± 0.02	0.87 ± 0.01
ADRA2A	0.90 ± 0.01	0.75 ± 0.01	0.84 ± 0.009	0.77 ± 0.02	0.61 ± 0.03	0.90 ± 0.006
Histamine H1	0.97 ± 0.003	0.87 ± 0.005	0.94 ± 0.007	0.87 ± 0.008	0.65 ± 0.02	0.84 ± 0.01
CCR5	0.92 ± 0.007	0.89 ± 0.008	0.90 ± 0.006	0.86 ± 0.01	0.68 ± 0.02	0.91 ± 0.009
hERG	0.83 ± 0.02	0.51 ± 0.01	0.79 ± 0.01	0.66 ± 0.02	0.60 ± 0.02	0.71 ± 0.01

Table 1. The AUC of our method, RMD, outperforms the random matrix theory benchmark [15] as well the naive Bayes, Tanimoto similarity, SVM and graph convolutional fingerprint methods.

can interpret the model to offer novel pharmacological insights about the roles of different chemical motifs in determining activity.

Our work significantly extends a random matrix framework developed by some of us [15] by having an explicit statistical model for the inactive set, prospective experiments on a therapeutically relevant receptor and robust retrospective tests with confirmed inactives, as well as offering new biochemical hypotheses through model visualisation and interpretation. To our knowledge this is the first *prospective* validation of the random matrix methodology applied to drug discovery. Our work also demonstrates the importance of high quality negative data and methodologies that clean and exploit negative correlations.

Random matrix theory and chemical correlations

We focus on a popular set of descriptors used in chemoinformatics. Molecular fingerprints are typically constructed by first representing a molecule as a 2D molecular graph, and then considering all possible bond paths (contiguous atoms connected by chemical bonds) within the molecule. Only identical molecules would share the same bond paths, and similar molecules share most bond paths, thus comparing bond paths is a reasonable way to quantify chemical similarity. As the set of all possible bond paths is vast, typically fingerprints are defined by first considering bond paths that are within some radius of every atom in the molecule, and then mapping these bond paths to a bit string of defined length through a hash function. Throughout this paper, we will use the 1024 bit Morgan 3 fingerprint [19] generated using the open-source package `rdkit` [28].

To determine which bond paths are important for binding, we need to determine which bond paths are correlated in their presence/absence in the set of active molecules relative to the set of inactive molecules. Principal component analysis provides way to do that. Mathematically, each molecule can be characterised as a binary vector of length p, $\mathbf{f}_i \in \mathbb{R}^p$. The ensemble of N_+ active molecules can be arranged as a data matrix $R_+ = [\mathbf{f}_1; \mathbf{f}_2 \cdots \mathbf{f}_{N_+}] \in \mathbb{R}^{N_+ \times p}$, and similarly R_- can be constructed from the N_- inactive molecules. We rescale the features such that the columns of $R_\pm$ have zero mean and unit variance; columns with zero variance are removed. Persistent correlations in bond paths can be identified from the eigendecomposition of each sample covariance matrix

$$C_\pm = \frac{1}{N_\pm} R_\pm^T R_\pm = \sum_{i=1}^{p} \lambda_i^\pm \mathbf{v}_i^\pm \otimes \mathbf{v}_i^\pm, \qquad [1]$$

where $\{\lambda_i^\pm\}$ are the eigenvalues and $\{\mathbf{v}_i^\pm\}$ are the eigenvectors. Each eigenvector $\mathbf{v}_i^\pm$ identifies a particular combination of the p bond paths which explains a fraction $\lambda_i / \sum_i \lambda_i$ of the variance.

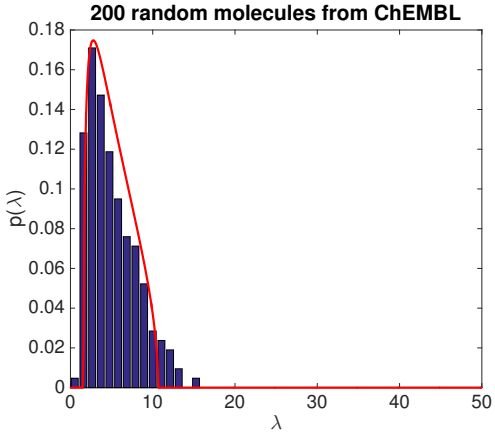

Fig. 1. The eigenvalue distribution of random molecules drawn from ChEMBL follows the random matrix distribution. The histogram shows the eigenvalue distribution of 200 random molecules drawn from ChEMBL, and the red curve is the random matrix distribution, Equation (2), for p = 1024 and N = 200.

However, not all eigenvectors are equally important. The question of discriminating signal from noise due to the undersampling in the context of molecular fingerprints has been discussed in the literature [15, 29, 30]. Under certain weak assumptions, if entries in $R_\pm$ are random and drawn from a Gaussian distribution with zero mean and unit variance, the probability of $C_\pm$ having an eigenvalue λ is given by the Marchenko–Pastur distribution [31]

$$\rho_\pm(\lambda) = \frac{\sqrt{\left[(1+\sqrt{\gamma_\pm})^2 - \lambda\right]_+^2 \left[\lambda - (1-\sqrt{\gamma_\pm})^2\right]_+^2}}{2\pi\gamma_\pm\lambda}, \qquad [2]$$

where $\gamma_\pm = p/N_\pm$ describes how well-sampled the dataset of active or inactive molecules is, and $(\cdot)_+ = \max\{\cdot, 0\}$. Equation (2) provides a suitable null distribution – only eigenvectors with eigenvalues outside the Marcenko–Pastur distribution are statistically significant. In practice, as the Marcenko–Pastur distribution is non-zero only in the region $\lambda \in [(1-\sqrt{\gamma_\pm})^2, (1+\sqrt{\gamma_\pm})^2]$, only eigenvectors with an eigenvalue greater than $(1+\sqrt{\gamma_\pm})^2$ are significant. In other words, for a less well-sampled dataset ($\gamma_\pm$ large), an eigenvector needs to have a large eigenvalue, i.e. explains a lot of the variance in the data, before one can believe that it is significant. The statistically significant orthonormal eigenvectors are orthogonal chemical features that are relevant for binding. If there are $m_\pm$ significant eigenvalues, then the linear space spanned by those $m_\pm$ associated eigenvectors, $\mathbf{V}_\pm = \text{span}\{\mathbf{v}_1^\pm, \mathbf{v}_2^\pm, \cdots \mathbf{v}_{m_\pm}^\pm\}$, is

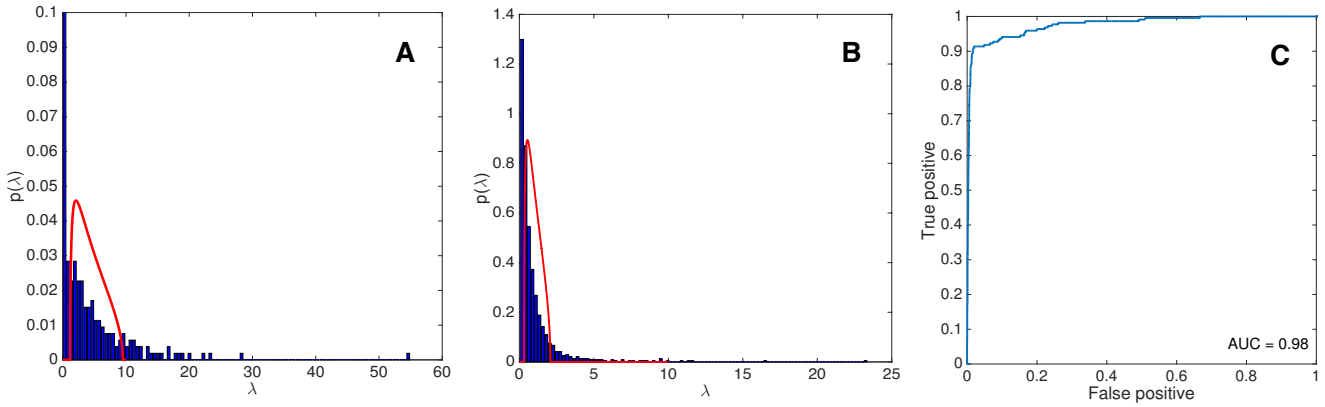

Fig. 2. Our random matrix model captures the statistics of M1 agonists and confirmed inactives from a historic campaign. The random matrix distribution (red curve) agrees with the histogram of eigenvalues of the (A) active agonists and (B) confirmed inactives. (C) A classification model built using the statistically significant eigenvectors achieves an accuracy of 98%.

the subspace of chemical feature space that causes binding/non-binding to that particular receptor.

Intuitively, a molecule is likely to be active if it is chemically similar to the set of known active molecules and dissimilar to the set of known inactive molecules. We can capture this intuition by requiring the molecule, represented as the Morgan fingerprint $\mathbf{u}$, to be close to the linear subspace $\mathbf{V}_+$ but far from $\mathbf{V}_-$. Therefore, we should classify a molecule $\mathbf{u}$ as active if

$$\mathcal{D}(\mathbf{u}, \mathbf{V}_+) < \mathcal{D}(\mathbf{u}, \mathbf{V}_-) + \epsilon, \qquad [3]$$

where ϵ is a tolerance parameter that captures the tradeoff between false positive and false negative, and $\mathcal{D}(\mathbf{u}, \mathbf{V}_\pm) = \left|\left|\mathbf{u} - \sum_{i=1}^{m_\pm} \left[\mathbf{v}_i^\pm \cdot (\mathbf{u} - \mu_\pm)/\sigma_\pm\right] \mathbf{v}_i^\pm\right|\right|_2$ is the distance between the vector $\mathbf{u}$, translated by the mean of the active/inactive set $\mu_\pm$ and scaled by the variance $\sigma_\pm$, to the linear subspace $\mathbf{V}_\pm$.

Equation (3) is the central result of this paper. It extends the result of our previous work [15] by explicitly accounting for the set of inactive molecules. This is important as the set of molecules tested are often clustered around scaffolds, and those scaffolds will appear as statistically significant eigenvectors regardless of whether they are important for binding. Therefore, one should focus only on correlations that are present in the active but *not* in the inactive set, which is the interpretation of Equation (3). Henceforth we will refer to Equation (3) as the Random Matrix Discriminant (RMD).

Retrospective benchmarks

We first validate the random matrix distribution as a suitable null hypothesis. Figure 1 shows that the eigenvalue distribution of 200 random molecules drawn from ChEMBL [32] indeed agrees with the random matrix null distribution, Equation (2). The Kolmogorov-Smirnov test statistic is $D = 0.087$, thus the hypothesis that the eigenvalue distribution follows Equation (2) cannot be rejected at the 0.95 confidence level. A small number of eigenvalues are outside the threshold predicted by Equation (2) because of two reasons: First, Equation (2) is derived asymptotically in the limit $p, n \to \infty$ with p/n fixed, thus there are finite size corrections that we neglected. Second, Equation (2) describes the typical behaviour of random matrices rather than extreme value distributions. [33] Nonetheless, by analysing 10 batches of 200 random molecules drawn from ChEMBL, we find that more than 95% of the total number of eigenvalues are within the bound predicted by Equation (2), thus Equation (2) is a suitable null hypothesis.

Following our previous study [15], we benchmark RMD using the challenge of identifying ligands of human G-Protein Coupled Receptors (GPCRs). We consider GPCRs where there are more than 500 known active molecules in ChEMBL and more than 150 inactive molecules from the internal Pfizer database; we consider only structures that are already in the public domain so that the dataset can be fully disclosed in the Supplementary Information. A ligand is considered active against a given target if its K_i, K_d, IC_{50}, or EC_{50} is 1 μM or less, and inactive otherwise. Our previous study [15] only considered active molecules because confirmed inactives from a "pharmacologically plausible" chemical space are difficult to obtain from the literature. This study importantly benefits from high quality negative data from proprietary historic high throughput screening campaigns. All in all, 4 receptors – μ opioid receptor (MOR1), serotonin receptor 2B (5-HT2B), α-2A adrenergic receptor (ADRA2A) and histamine H1 receptor – have sufficient number of disclosable confirmed inactives.

Table 1 shows that RMD outperforms the benchmark [15] as well as the naive Bayes method and classification based on the mean Tanimonto coefficient against active molecules in the training set. The latter two methods are common chemoinformatics methods used in the industry. Our method takes into account the mean and pairwise correlations of chemical features in a noise-robust manner, thus a natural question to ask is whether our method outperforms non-linear methods in the literature that use higher order correlations. To make this comparison, we focus on the Support Vector Machine (SVM) with a cubic kernel (which accounts for third order correlations) and the Graph Convolutional Neural Network fingerprint [26]. Table 1 also shows that RMD outperforms SVM as well as the Graph Convolutions (implemented in the open-source package `DeepChem` [34]). In all tests, the active and inactive sets are randomly split into a training set (90%) and test set (10%). The figure of merit that we consider is the area under curve (AUC) of the receiver operating characteristic. The mean AUC and standard error of the mean are estimated by analysing 10 random partitions. In the Supplemental Material we show that our method also outperforms other common machine learning

methods for the Avalon fingerprint, another descriptor based on a handcrafted set of chemical features [17], showing that the importance of cleaning undersampling noise is independent of descriptor choice. The AUCs of the two fingerprints with RMD are comparable, and we use the Morgan 3 fingerprint for ease of directly interpreting the model in terms of correlations between chemical fragments, c.f. Figure 4.

One looming question is whether RMD performs well because the chemical space probed by Pfizer is different to the published literature on ChEMBL. To answer this question, we search our database for a receptor system where there are more that 500 active molecules and 150 inactive molecules. The receptor system that fulfils those criteria is the chemokine receptor type 5 (CCR5). Table 1 also shows that RMD works equally well when the chemical space of both the active and inactive molecules originate from the same source.

Thus far we have only considered GPCRs. However RMD is agnostic as to whether the target is a GPCR or even a protein receptor. To illustrate this point, Table 1 shows that RMD can accurately predict binding to hERG, a potassium ion channel that is an "anti-target" for drug discovery as binding can cause cardiac arrest [35, 36]. Both the active and inactive data come from internal screens. Following commonly used thresholds for hERG activity [35, 36], we classify compounds with an $\mathrm{IC}_{50} < 9\mu M$ as active and $\mathrm{IC}_{50} > 29.9\mu M$ as inactive.

Prospective discovery of novel human M1 agonists

We select the human muscarinic acetylcholine receptor M1 as a target to prospectively deploy our algorithm. The muscarinic acetylcholine receptors are members of the rhodopsin-like GPCRs and regulate the functions of the central and peripheral nervous system. Currently, drugs that target muscarinic receptors have been approved for the treatment of chronic obstructive pulmonary disease, overactive bladder and Sjogren's syndrome [37, 38]. Gene knockout studies suggest that the M1 agonists may ameliorate the symptoms of Alzheimer's disease and related cognitive disorders, as well as ameliorate psychosis-like symptoms in Schizophrenia [37, 38].

We use data from a historic campaign to train our model. In total, 5445 compounds were screened for agonist activity, out of which 222 were active ($\mathrm{EC}_{50} < 1\mu M$). It is apparent that the number of molecules, even in an industrial campaign, is small compared to the number of chemical features, thus removing undersampling noise is essential. Figure 2A-B shows that the eigenvalue distribution of the active and inactive set follows Equation (2), save for a few statistically significant eigenvectors. Figure 2C reassures the reader by showing the classification accuracy of RMD on the historic data; as before the data is partitioned into training/testing (90%/10%) sets.

The model is then deployed to screen the entire `e-Molecules` database [39], a publicly accessible database of 5.9 million commercially available chemical compounds. We select the top 150 molecules using RMD, the naive Bayes classifier as well as by maximum Tanimoto similarity to training set, and perform a prospective experimental test (see Methods). Figure 3 shows that RMD discovered 4 novel agonists. The closest structure in the training set by Tanimoto similarity is shown in the insets of Figure 3. RMD has found active molecules that are structurally dissimilar to molecules in the training set, and can successfully hop between chemical scaffolds. In terms of Tanimoto similarity, none of the new agonist

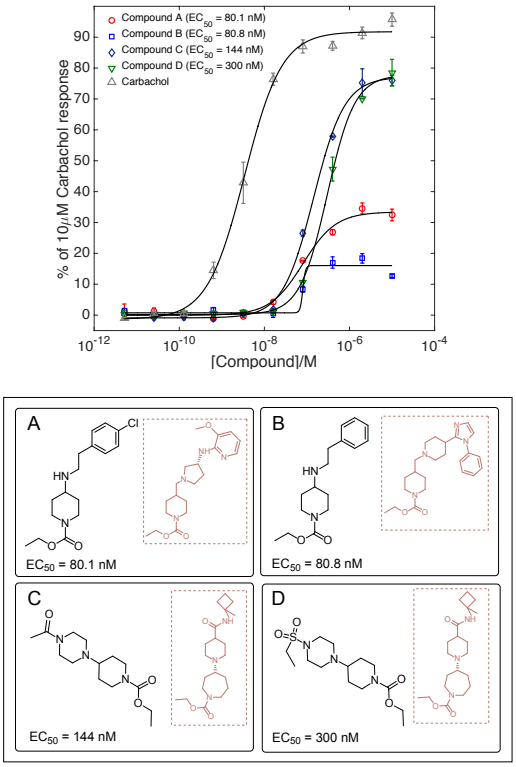

Fig. 3. The RMD model discovered 4 novel human M1 agonists. Top panel: the measured dose-response curves for the novel agonists. Bottom panel: The molecular structures of the agonists; the insets show the closest molecule in the training set by Tanimoto coefficient.

we found has a Tanimoto coefficient greater than 0.41 to any molecule in the training set – this level of (dis)similarity is the same as two molecules randomly drawn from ZINC, a large database often used in virtual screening [40].

The hit rate of RMD is also greater than common chemoinformatics methods. For comparison, the naive Bayes classifier only discovered 2 novel agonists and the agonists predicted by the Tanimoto classifier are all inactive. Moreover, the hit rate of a typical high throughput screen is $\sim 0.01\% - 0.14\%$ [41], thus our model performance is around almost 40-fold better than the background hit rate. This finding corroborates the results from the retrospective test, Table 1.

Model interpretation and visualisation

The ability of RMD to prospectively discover novel M1 agonists that are chemically dissimilar to the training set motivates us to unbox the model and interpret its "reasoning". RMD classifies molecules based on their distance from the linear subspace spanned by motifs in the active set relative to the inactive set. As such, an intuitive way to interpret the model is to directly visualise the difference between the two linear subspaces. We can operationalise this difference by defining the following effective correlation matrix

$$\mathcal{A} = \sum_{i=1}^{m_+} \mathbf{v}_i^+ \otimes \mathbf{v}_i^+ - \sum_{i=1}^{m_-} \mathbf{v}_i^- \otimes \mathbf{v}_i^- . \qquad [4]$$

A positive entry $\mathcal{A}_{ij}$ denotes that motifs i and j are jointly and positively contributing to binding, whereas a negative entry

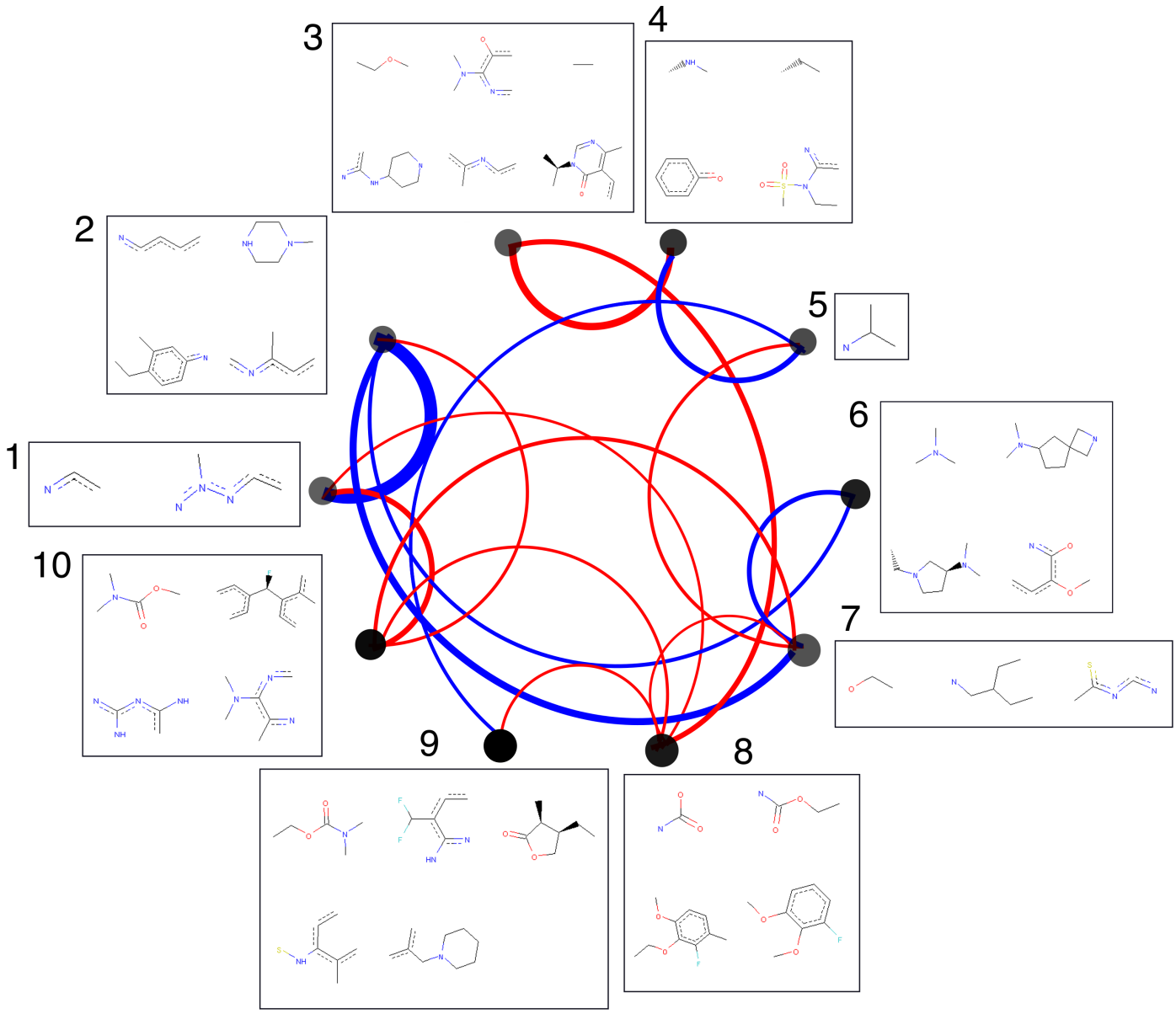

Fig. 4. Our model can be interpreted as a network of features, where each feature is an entry of the molecular fingerprint. The opacity of the nodes is proportional to the difference in the number of times that feature is present in the active set relative to the inactive set; only the top 10 features are shown. Red (blue) edges correspond to a positive (negative) correlation, and the width of the edges is proportional to the strength of the correlation.

denotes that the motifs are jointly and negatively effecting binding. A similar approach have been proposed in the context of principal component analysis where the goal is to reveal the contrast between two groups [42].

Figure 4 shows that the motif-motif correlation matrix (4) can be visualised and interpreted as a network of chemical fragments. Each motif is an entry of the molecular fingerprint. Due to bit collision, multiple fragments can be assigned to the same motif. Interestingly, the model identifies the carbamate fragment (node 10 in Figure 4) as distinct and positively correlated to the piperazine fragments (node 2), and puts them together to form the new active molecules C and D (c.f. Figure 3). Crucially, the model is able to learn that a carbamate with a ternary nitrogen is the relevant fragment, rather than a 6-membered ring piperidine-N-carboxylate or 7-membered ring azepane-N-carboxylate present in the training set. In other words, the model interpolates between structures in the training set by learning generalisable chemical features rather than simply memorising the training data. Another trend that the model extracts is the correlation between carbamate (node 10) and aromatic fragments (nodes 1 and 2), agreeing with the heuristic in the literature that M1 agonists generally have a hydrogen bond acceptor and a distal aromatic moiety [43].

The negative correlations are also chemically significant. One interesting prediction is that although the aromatic nitrogen fragment (node 1) and the piperazine fragment (node 2) are both overrepresented in the active molecules, having *both* fragments is strongly detrimental to agonist activity. Intriguingly, the model does not predict a similar negative correlation between the aromatic nitrogen fragment and the cognate piperidine fragment, which is structurally identical to piperazine except having one less nitrogen atom. This subtle prediction of an activity difference between piperidine and piperazine is ripe for experimental testing.

We search our internal database for pairs of molecules that are neither used in model training nor testing nor in the $\mathtt{e-Molecules}$ database, both containing an aromatic nitrogen moiety, and are exactly the same except one has a piperidine ring and the other has a piperazine ring. Those "matched molecular pairs" [44] were experimentally tested for M1 agonist activity. Note that none of those matched molecular pairs are considered in the original model, thus this is a completely independent out-of-sample validation of model prediction. Figure 5 shows that in all cases, swapping the piperazine motif for the piperidine motif leads to a significant increase in activity. This confirms the strongly negative motif-motif correlation that our model has picked up as well as demonstrating how we can exploit this chemical insight to introduce a small chemical modification to the molecule that significantly increases binding affinity. The visualisation in Figure 4, combined with insights on synthetic accessibility, can provide ideas for de novo design.

Conclusion

We derived a statistical framework inspired by random matrix theory for ligand biological activity prediction that discovers important correlations between chemical features by disentangling undersampling noise and subtracting the correlation structure of the active compounds from the inactive compounds. We showed that the model outperforms standard benchmarks when tested against a set of challenging retrospective tests. We prospectively applied the model to the human muscarinic acetylcholine receptor M1 and found 4 experimentally-confirmed novel agonists that are chemically dissimilar to all known agonists. Moreover, we can visualise and interpret the model to yield novel pharmacological insights. Our method distills which combinations of chemical motifs are positively/negatively responsible for binding to the receptor, and predicts pairs of motifs of which each individual motif is overrepresented in the active molecule (thus naively expected to be important for binding) yet when occurred as a pair are detrimental to binding. We experimentally validated the pharmacological insights predicted by the model, showing how one can exploit insights afforded by the model to make a small chemical change to a molecule to evade those negative motif pairs and drastically improving potency. A broader conclusion of our study is the power of high-quality inactive data, which allows the model to generate meaningful hypothesis about motif combinations that lead to inactivity.

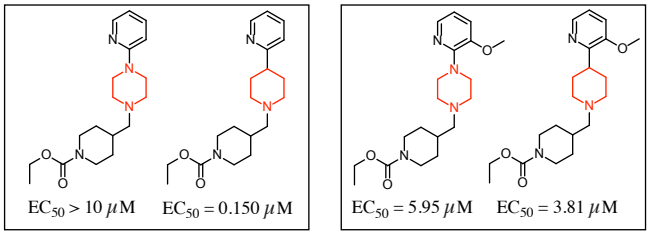

Fig. 5. Prospective matched molecular pair analysis corroborates the significant negative correlation between the piperazine and the the aromatic nitrogen motif that the model predicts.

Methods

M1 agonist assay. Chinese hamster ovary (CHO) cells stably expressing the human muscarinic 1 receptor were plated at a density of 7,500 cells per well (50 μL per well) in black walled clear bottomed 384-well plates and were incubated overnight (20-24 hours) in a 37°C humidified incubator with 5% carbon dioxide. Agonist activity was determined by measuring compound stimulated changes in intracellular calcium levels using a calcium sensitive dye. Prior to the start of the experiment the medium was removed from the plates and 80 μL of Hanks balanced salt solution (HBSS) containing HEPES (20 mM), Calcium 5 dye (Molecular Devices, Sunnyvale, CA; Catalogue Number R8186) and probenecid (1.25 mM) was added to each well and the plate was returned to the incubator for 1 hour to allow dye loading. For compound assessment in the agonist format 10 μL of the compound solution was added to each well by the FLIPR Tetra instrument (Molecular Devices, Sunnyvale, CA) and the change in fluorescence from baseline over a 60 second period (Ex 470-495 nM; Em 515-575 nM) was measured.

Data and code availability. Active and inactive compounds for MOR1, 5-HT2B, ADRA2A, histamine H1, and hERG, as well as data from the prospective experiments on M1, are available in the SI. The ode used to perform RMD analysis is on GitHub (`github.com/alphaleegroup/RandomMatrixDiscriminant`).

ACKNOWLEDGMENTS. AAL acknowledges support from the Winton Programme for the Physics of Sustainability.

1. Alvarez J, Shoichet B, eds. (2005) *Virtual screening in drug discovery*. (CRC press).
2. Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug discovery today* 11(13-14):580–594.
3. Kubinyi H, Mannhold R, Timmerman H, et al. (2008) *Virtual screening for bioactive molecules*. (John Wiley & Sons) Vol. 10.
4. Koeppen H (2009) Virtual screening-what does it give us? *Current opinion in drug discovery & development* 12(3):397–407.
5. Schneider G (2010) Virtual screening: an endless staircase? *Nature Reviews Drug Discovery* 9(4):273.
6. Lyne PD (2002) Structure-based virtual screening: an overview. *Drug discovery today* 7(20):1047–1055.
7. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH (2012) Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS journal* 14(1):133–141.
8. Lionta E, Spyrou G, K Vassilatis D, Cournia Z (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry* 14(16):1923–1938.
9. Chodera JD et al. (2011) Alchemical free energy methods for drug discovery: progress and challenges. *Current opinion in structural biology* 21(2):150–160.
10. Hansen N, Van Gunsteren WF (2014) Practical aspects of free-energy calculations: a review. *Journal of chemical theory and computation* 10(7):2632–2647.
11. Geppert H, Vogt M, Bajorath Jr (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of chemical information and modeling* 50(2):205–216.
12. Ripphausen P, Nisius B, Bajorath J (2011) State-of-the-art in ligand-based virtual screening. *Drug discovery today* 16(9-10):372–376.
13. Keiser MJ et al. (2007) Relating protein pharmacology by ligand chemistry. *Nature biotechnology* 25(2):197.
14. Keiser MJ et al. (2009) Predicting new molecular targets for known drugs. *Nature* 462(7270):175.
15. Lee AA, Brenner MP, Colwell LJ (2016) Predicting protein–ligand affinity with a random matrix framework. *Proceedings of the National Academy of Sciences* 113(48):13564–13569.
16. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences* 42(6):1273–1280.
17. Gedeck P, Rohde B, Bartels C (2006) Qsar- how good is it in practice? comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of chemical information and modeling* 46(5):1924–1936.
18. Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics* 5(1):26.
19. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50(5):742–754.
20. Schneider N, Fechner N, Landrum GA, Stiefl N (2017) Chemical topic modeling: Exploring molecular data sets using a common text-mining approach. *Journal of chemical information and modeling* 57(8):1816–1831.
21. Laloux L, Cizeau P, Bouchaud JP, Potters M (1999) Noise dressing of financial correlation matrices. *Physical review letters* 83(7):1467.
22. Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Stanley HE (1999) Universal and nonuniversal properties of cross correlations in financial time series. *Physical review letters* 83(7):1471.
23. Bun J, Bouchaud JP, Potters M (2017) Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports* 666:1–109.
24. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2009) The graph neural network model. *IEEE Transactions on Neural Networks* 20(1):61–80.
25. Lusci A, Pollastri G, Baldi P (2013) Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling* 53(7):1563–1575.
26. Duvenaud DK et al. (2015) *Convolutional networks on graphs for learning molecular fingerprints*. pp. 2224–2232.
27. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P (2016) Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* 30(8):595–608.
28. (year?) RDKit: Open-source cheminformatics (http://www.rdkit.org).
29. Lee AA, Brenner MP, Colwell LJ (2017) Optimal design of experiments by combining coarse and fine measurements. *Physical review letters* 119(20):208101.
30. Cortes Cabrera A, Petrone PM (2018) Optimal hts fingerprint definitions by using a desirability function and a genetic algorithm. *Journal of chemical information and modeling*.
31. Marčenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* 1(4):457.
32. Gaulton A et al. (2016) The chembl database in 2017. *Nucleic acids research* 45(D1):D945–D954.
33. Johnstone IM (2001) On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics* pp. 295–327.
34. Wu Z et al. (2018) Moleculenet: a benchmark for molecular machine learning. *Chemical Science* 9(2):513–530.
35. C Braga R et al. (2014) Tuning herg out: antitarget qsar models for drug development. *Current topics in medicinal chemistry* 14(11):1399–1415.
36. Didziapetris R, Lanevskij K (2016) Compilation and physicochemical classification analysis of a diverse herg inhibition database. *Journal of computer-aided molecular design* 30(12):1175–1188.
37. Wess J, Eglen RM, Gautam D (2007) Muscarinic acetylcholine receptors: mutant mice provide new insights for drug development. *Nature reviews Drug discovery* 6(9):721.
38. Kruse AC et al. (2014) Muscarinic acetylcholine receptors: novel opportunities for drug development. *Nature reviews Drug discovery* 13(7):549.
39. (year?) e-molecules (https://www.emolecules.com).
40. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2013) Molecular similarity in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* 57(8):3186–3204.
41. Zhu T et al. (2013) Hit identification and optimization in virtual screening: practical recommendations based on a critical literature analysis: miniperspective. *Journal of medicinal chemistry* 56(17):6560–6572.
42. Abid A, Bagaria VK, Zhang MJ, Zou J (2017) Contrastive principal component analysis. *arXiv preprint arXiv:1709.06716*.
43. Bhandare RR, Canney DJ (2011) Modifications to five-substituted 3, 3-diethyl-4, 5-dihydro-2 (3h)-furanones en route to novel muscarinic receptor ligands. *Medicinal Chemistry Research* 20(5):558–565.
44. Griffen E, Leach AG, Robb GR, Warner DJ (2011) Matched molecular pairs as a medicinal chemistry tool: miniperspective. *Journal of medicinal chemistry* 54(22):7739–7750.